

PATENT APPLICATION

for

5

METHOD AND APPARATUS FOR COMPILING BUSINESS DATA

By

Jeffrey K. Wilkins and Prashant Devdhar

CROSS-REFERENCE TO RELATED APPLICATIONS

This application is cross-referenced to and claims priority from U.S Provisional application 60/219,751 filed July 20th, 2000, which is hereby incorporated by reference.

FIELD OF THE INVENTION

20 The present invention relates generally to computer systems. More particularly, the present invention relates to an Internet-based system for identifying and compiling potential business data based on Internet-accessible data.

BACKGROUND

According to the Direct Marketing Association, around \$100 billion was spent on US business-to-business direct marketing at the beginning of the new millennium. Of this total, telemarketing accounts for about 40%, and direct mail expenditures total about 16%.

5 Business marketers typically target prospects within a business based on their ability to recommend, influence or approve a purchase. Commonly, the marketer obtains, for example, a list of the name, title, postal address, and/or telephone numbers of these prospective buyers so that direct marketing programs can be initiated.

10 Further elaborating on this example, typically, a prospect must have the ability, willingness, and readiness to buy before a direct sale can be consummated. One method is to identify individuals with certain titles (e.g. CEO, VP Marketing) that are likely to correlate with purchasing authority. Vendors, such as Dun & Bradstreet, Murray Hill, NJ (www.dnb.com) and InfoUSA, Omaha, NE (www.infousa.com), gather business data through a variety of means. For example, the sources for InfoBase Business List (Acxiom Corporation, Little Rock, AR, www.acxiom.com) include:

- Annual reports, 10Ks, and other SEC information. These sources provide accurate, in-depth information about publicly traded companies including the name of executives and directors, employment figures, sales volume, and more.
- Federal, state and municipal government filings. These sources provide current directories, reports, and Chamber of Commerce information.
- Business magazines, newsletters and newspapers. These sources provide news about business merger, acquisitions, executive changes and financial results.
- Outbound telemarketing. These sources provide information about small businesses and their owners obtained from firms who frequently market to these businesses.

- Professional directories. These sources provide invaluable information about professionals, their clinics and partnerships, and their specialties.

Leveraging these sources, the InfoBase business list provides over 14 million business names
5 and addresses representing over 12.9 million separate businesses.

There are several limitations in utilizing these sources for compiling, for instance, a list of senior managers. The limitations include:

- Recency. American society has been characterized by shorter tenures of employment and greater job mobility. But many of the sources used to compile lists of senior managers are updated infrequently. In some instances, data is refreshed only once every 1-2 years. This means that changes in senior management, such as new hires and terminations, go unreported for a significant time.
- Cost. Because it is so labor intensive, it is often prohibitively expensive to conduct telephone or face-to-face interviews, or verify contact information 'on demand'.
- Comprehensiveness. Often, only a small number of contacts are gathered at each business. For example, Acxiom's list contains an average of 1.17 contacts per company (14 million contact names at 12 million US businesses). What's more, contact data for non-US companies is even scarcer.

Accordingly, there is a need to develop a more useful and efficient approach to compile
25 business data that overcomes the limitations in the prior art solutions.

SUMMARY OF THE INVENTION

The present invention provides a method and apparatus of compiling and providing business data in a distributed computer system, preferably the Internet or any other type of computer linked network system. The present invention also provides a computer program product that contains business data compiled from files containing business data located in a distributed computer system. The product according to the present invention could be stored in any type of computer readable medium that stores business data. The present invention includes locating files or links in a distributed computer system that contain business data. Once the files or links have been identified and selected, each file or link is parsed to extract the business data of interest. After the business data of interest is extracted, the extracted business data could be transferred to an interested party. The present invention also includes different ways of evaluating the files or links that contain potential business data. The evaluation includes the determination of a confidence level of finding the business data. However, the evaluation also includes the determination of a confidence level of finding a subset of the business data. In addition, the evaluation includes the use of keywords in the file or link to determine a confidence level of finding the business data. The files or links are usually located on, but not limited to, corporate, magazine, newspaper, press release, professional, and association websites. The location of files or links is accomplished by different ways, such as, but not limited to, using a publicly accessible search engine, using a general purpose or custom defined spider, or by selecting one or more links in the computer distributed system wherein the links are selected based on their proximity to a set of keywords. The location of the files or links could also be established by using a list that defines the files or links based on a previously generated list of the files. The present invention further includes the step of evaluating a tense, for instance a past or a future tense, related to the business data. Furthermore, once the tense is identified the present invention

determines whether or not to include or discard the business data that is associated with that particular tense. The present invention provides mechanisms to use one or more tags to locate the files or links that contain business data. In addition, the present invention provides mechanisms to use one or more tags to extract the business data. Furthermore, the present invention includes the step of creating a concordance table to facilitate the extraction of the business data. For instance, in the concordance table one or more tag counts are associated with the business data. In addition, the present invention includes a mechanism to rate and bound the local content group that contain the business data of interest. The present invention also provides an analysis that uses inclusion and exclusion characteristics to extract the business data. The present invention further includes the step of normalizing the business data. Eventually, the present invention also enables one to eliminate duplicate sets of business data. In order to keep the extracted business data current, the present invention includes the step of extracting date or time stamps of the files that contain the business data. In the updating process, the date or time stamps are compared with the date or time stamps in the previously extracted business data or in the located files themselves. Date or time stamps are also used to identify particular business data before extraction.

In view of that which is stated above, it is the objective of the present invention to provide a cost-effective method for automatically compiling a list of large numbers of business data.

It is another objective of the present invention to provide an automated method for refreshing business data so as to keep it current.

It is yet another objective of the present invention to provide a method that takes advantage of the large amounts of business data that is available on the Internet and that is constantly updated.

It is still another objective of the present invention to provide a method and apparatus for generating lists of business data that is highly scalable, both in the number and their geographic distribution.

5

The advantage of the present invention over the prior art is that the system enables one to compile business data from files in a distributed computer system in a cost effective and time effective manner taking advantage of the large number of data that is available in a distributed computer system on the Internet. Another advantage is that the present invention enables one to quickly update business data to stay current.

10
15
20
25

BRIEF DESCRIPTION OF THE FIGURES

The objectives and advantages of the present invention will be understood by reading the following detailed description in conjunction with the drawings, in which:

FIG. 1 shows an illustrative diagram of a method according to the present invention; and
FIG. 2 shows an exemplary flowchart of steps for processing files or links that contain business data according to the present invention.

20

DETAILED DESCRIPTION OF THE INVENTION

Although the following detailed description contains many specifics for the purposes of illustration, anyone of ordinary skill in the art will readily appreciate that many variations and alterations to the following exemplary details are within the scope of the invention.

25 Accordingly, the following preferred embodiment of the invention is set forth without any loss of generality to, and without imposing limitations upon, the claimed invention.

The present invention provides a method of providing business data in a distributed computer system, preferably the Internet or other types of computer linked network systems. The present invention is used to identify and compile business data that relates to any type of business data that is entertained in a business and/or businesses. The present invention does not include any type of business data whether it is from a non-profit, for-profit or government organization. An example of the scope of the present invention is, for instance, to identify websites conducting e-commerce. For this particular application one would use the presence of shopping carts, check-out pages, credit card order forms, secure server https, etc. on various websites as an indicator to identify and compile business data. Another example of the scope of the present invention is extracting corporate contact information (postal address, telephone number, fax number) from websites. Naturally, this is useful to pair with the names of senior executives that one could extract to generate complete direct mail or telemarketing records. For this particular example, it also helps to tie the domain name of online businesses to business marketing databases such as D&B or infoUSA. It can also help identify 'corporate linkages' between offices or subsidiaries of the same company. Yet another example of the scope of the present invention is compiling e-mail lists by extracting the name and e-mail address of employees. As one skilled in the art would readily acknowledge, the present invention could compile any type of data from a computer distributed system and is not restricted to these examples.

The following provides an exemplary embodiment wherein the present invention is used to identify contact information for large numbers of senior managers in a cost effective manner. Contact information in this example encompasses name and job title, but in some instances also includes postal address, telephone number, email address or fax number. The present invention relies on the wealth of continuously updated business data available on the Internet

that can be mined to identify contact information for senior managers. Once identified, the contact information can be extracted and enhanced with direct marketing databases to allow further targeting.

5 The method of the present invention leverages business data from an employer's own corporate website, as well as third party sites. However, the present invention is not limited to any source in the computer distributed system that contains any relevant files that contain business data.

10 **Corporate Websites.** According to Internet registrar Network Solutions, approximately 14 million domain names were registered as of March 31, 2000, an increase of 233% from the 4.2 million registered as of March 31, 1999. It is estimated that fully 3.6 million or 26% are used by US businesses as their primary corporate website.

15 A corporation's own website is a prime information source for identifying the names, job titles, and other contact information for its senior managers. For example, many company websites contain a webpage listing the top 5-10 senior executives. The page minimally contains the name and job title of the managers. Sometimes this information is augmented with a bio of the manager, their email address, phone number, fax number or postal address.

20 For example, on July 10, 2000, the website for Lowell General Hospital, Lowell, MA, listed members of senior management on <http://www.lowellgeneral.org/HTML/Mgmt.html>: In the following example below, each manager's name was hyperlinked (indicated by an underline) to a detailed biography.

Senior Management at Lowell General
Management Profiles

5

Robert A. Donovan
President and Chief Executive Officer

10

Normand E. Deschene
Executive Vice President
and Chief Operating Officer

15

Nicholas J. Arloro
Senior Vice President, Finance
Treasurer

20

M. Patricia Crane
Vice President, External Affairs

25

Gail LaBrecque
Vice President, Operations
Chief Nursing Officer

30

Wayne E. Pasanen, M.D.
Vice President, Medical Affairs

Ellen Caille
Vice President, Operations

Lowell General Hospital, 295 Varnum Avenue, Lowell, Massachusetts 01854
Main Telephone: (978) 937-6000

35

The management page in the example above is a particularly good source of fresh data since corporations maintain the info and have a strong incentive to update the page when new managers are hired or employment is terminated. In addition, once a 'map' detailing the location of this page has been built, it is easy to check whether it has been updated on demand.

40

In some cases, a management contact page is not provided. In these instances, other portions of the corporate website may be mined such as a listing of press releases.

Corporate websites are an excellent source of press releases, since many of these releases are not distributed to wire services or appear elsewhere on the web (see 'Third Party Sites' below). In these instances, press releases can be mined only from the corporate site. Press releases are commonly found on pages with names like 'Press', 'News' or the like. Beyond the name of the web pages, a press release can usually be identified by its structure. For example, the body of a press release starts with the city and date of the release as indicated in the example below. The body of the press release commonly ends with a standard boilerplate description of the firm. Press releases often quote senior executives from one or more organizations. Once identified, the name and title of the executive can be easily parsed.

For example, Buck & Pulley, a Rochester, NY marketing communications firm, announced the signing of Veramark as a new client in a press release (<http://www.pulley.com/Press/veramark/veramark.htm>). In this particular example, which is provided below, Dan Foster, president and COO, Buck & Pulley, can be identified as a senior manager of the firm as of June 6, 2000.

Veramark Selects Buck & Pulley for Public Relations and Collateral Design Services

ROCHESTER, N.Y., June 6, 2000 – Veramark® Technologies, Inc. [NASDAQ: VERA], an international developer of Internet resource management and enterprise telemanagement products, has selected Buck & Pulley, Inc. to perform marketing communications planning, public relations, and sales collateral design services on its behalf. These activities will initially focus on VeraWeb, a new Internet accounting and resource management product. Anticipated billings were not disclosed.

"Veramark complements our growing base of telecommunications and IT clients, including N-Soft and Corning Semiconductor Materials," said Dan Foster, president and COO, Buck & Pulley. "While the agency is expanding to serve technology clients throughout the U.S. and overseas, it's always a pleasure to work with a locally-based company."

5 Headquartered in Pittsford, N.Y., Veramark Technologies, Inc. is engaged in the design, production, and servicing of Internet, communications management, and billing systems for users and providers of network services worldwide. For additional information, visit www.veramark.com.

10 Founded in 1982, Buck & Pulley - a full service marketing communications firm based in Rochester, N.Y. - is a partner in the WORLDCOM Public Relations Group and a member of the American Association of Advertising Agencies. Its clients include Eastman Kodak Company, Xerox Corporation, Corning Inc., N-Soft S.A., National Industries for the Blind, Harris Interactive, and eBidenergy.com.

15

For more information, call (716) 288-6900.

20 **Third Party Websites.** In some cases, a company will not have a corporate website, or the website may not list senior executives or provide press releases. In these instances, third party websites may be useful. Many sources are web-accessible, including magazine and newspaper sites, press release wire service websites, or professional/trade association websites.

25 **Magazines/Newspapers.** More than 1,000 US newspapers are now Internet-accessible. Many are general interest daily newspapers, which include business sections containing both local and national news. Business data/contact information can also be found online in 'local' business newspapers (e.g. Austin Business Journal, Austin, Texas), national business publications (e.g. Businessweek) and trade magazines (e.g. EE Times) are also online. In some cases, this data can be accessed through aggregator websites which allow full-text searching of numerous sources. For example, Infonautics eLibrary (www.infonautics.com) provides access to more than 7.5 million articles from more than 150 full-text newspapers, hundreds of magazines, two international newswires, and more. Bizjournals.com provides a

searchable archive of its 41 local city business journals. In some instances, data must be accessed from an individual site. The site may provide a search engine for its own site, or a third party indexer/search engine such as Google or AltaVista may be used.

5 **Press Release Wire Services.** Press releases can also be found at wire service website companies such as BusinessWire (www.bizwire.com), PRNewswire (www.prnewswire.com), or financial portals like Yahoo! Finance (for example, Businesswire press releases can be found at <http://biz.yahoo.com/bw>). On July 10th, 2000, hyperlinks to the full text of numerous releases were listed, including those shown in the following example.

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

Business Wire - Monday July 10, 2000

(all times are Eastern)

[General Magic Appoints Paula E. Skokowski to Vice President of Marketing](#) - 9:04 am
[FiberChem Reports Successful E-Warrant Redemption](#) - 9:04 am
[eProNet Partners With the Forum for Women Entrepreneurs to Provide FWE Members Access to Top-Tier Job Candidates](#) - 9:04 am
[Accelerated Networks Joins Siemens to Demonstrate Multiservice Broadband Wireless Solution at WCA 2000](#) - 9:04 am
[Prodeo Technologies Announces New Pre-Owned Equipment Division, Prodeo XS](#) - 9:03 am
[Sensormatic Announces Change in CFO Position](#) - 9:03 am
[Road Runner Chooses BroadJump to Cut Time and Expense of Broadband Subscriber Support](#) - 9:03 am
[Mexico Must Resolve Issues for Top Investment Rating, Says Dupont COO In Energy Houston](#) - 9:03 am
[Percussion Software Sets Its Sights On the XML Content Management Market](#) - 9:03 am
[At Mid-Year, DVD-Video On Target to Beat Earlier Projections](#) - 9:02 am
30 [ADVISORY/Leading Registrar Namesecure to Provide Daily Notes and Commentaries From Japan ICANN Conference At \[www.namesecure.com\]\(http://www.namesecure.com\)](#) - 9:02 am
[I-Network.com Expands Sales Force to Meet Growing Demand for Services](#) - 9:02 am
[Alcoa's Earnings Up 57% From Year-Ago Quarter On a 38% Increase in Revenues](#) - 9:02 am
[Gulistan Carpet Covers Supply Chain Needs With Adexa](#) - 9:01 am

35

Professional/Trade Associations. Senior executives may be listed on trade/professional association websites. For example, as of July 10, 2000, the corporate website for the Auburn

Regional Medical Center, Auburn, WA (www.armcuhs.com) did not list any senior hospital administrators. But the names and titles of several executives were listed on the website of the Washington State Hospital Association (www.wsha.org/hospital/2000dir.htm). An extract of the listing for ARMC is shown below and lists the CEO, COO and business development administrator for the medical center.

5

Auburn Regional Medical Center
www.armcuhs.com
Plaza One, 202 N. Division Street - Zip 98001-4908
Michael M. Gherardini, Chief Executive Officer and Managing
Director, 333-2501
Pat Bailey, Assistant Admin., Business Development, 333-2528
D. Blain Claypool, Chief Operating Officer, 804-2860
General Hospital, 149 beds
Owned by Universal Health Services, Inc.
FAX: (253) 939-2376 (253) 833-7711

10
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
2035
2040
2045
2050
2055
2060
2065
2070
2075
2080
2085
2090
2095
2100
2105
2110
2115
2120
2125
2130
2135
2140
2145
2150
2155
2160
2165
2170
2175
2180
2185
2190
2195
2200
2205
2210
2215
2220
2225
2230
2235
2240
2245
2250
2255
2260
2265
2270
2275
2280
2285
2290
2295
2300
2305
2310
2315
2320
2325
2330
2335
2340
2345
2350
2355
2360
2365
2370
2375
2380
2385
2390
2395
2400
2405
2410
2415
2420
2425
2430
2435
2440
2445
2450
2455
2460
2465
2470
2475
2480
2485
2490
2495
2500
2505
2510
2515
2520
2525
2530
2535
2540
2545
2550
2555
2560
2565
2570
2575
2580
2585
2590
2595
2600
2605
2610
2615
2620
2625
2630
2635
2640
2645
2650
2655
2660
2665
2670
2675
2680
2685
2690
2695
2700
2705
2710
2715
2720
2725
2730
2735
2740
2745
2750
2755
2760
2765
2770
2775
2780
2785
2790
2795
2800
2805
2810
2815
2820
2825
2830
2835
2840
2845
2850
2855
2860
2865
2870
2875
2880
2885
2890
2895
2900
2905
2910
2915
2920
2925
2930
2935
2940
2945
2950
2955
2960
2965
2970
2975
2980
2985
2990
2995
3000
3005
3010
3015
3020
3025
3030
3035
3040
3045
3050
3055
3060
3065
3070
3075
3080
3085
3090
3095
3100
3105
3110
3115
3120
3125
3130
3135
3140
3145
3150
3155
3160
3165
3170
3175
3180
3185
3190
3195
3200
3205
3210
3215
3220
3225
3230
3235
3240
3245
3250
3255
3260
3265
3270
3275
3280
3285
3290
3295
3300
3305
3310
3315
3320
3325
3330
3335
3340
3345
3350
3355
3360
3365
3370
3375
3380
3385
3390
3395
3400
3405
3410
3415
3420
3425
3430
3435
3440
3445
3450
3455
3460
3465
3470
3475
3480
3485
3490
3495
3500
3505
3510
3515
3520
3525
3530
3535
3540
3545
3550
3555
3560
3565
3570
3575
3580
3585
3590
3595
3600
3605
3610
3615
3620
3625
3630
3635
3640
3645
3650
3655
3660
3665
3670
3675
3680
3685
3690
3695
3700
3705
3710
3715
3720
3725
3730
3735
3740
3745
3750
3755
3760
3765
3770
3775
3780
3785
3790
3795
3800
3805
3810
3815
3820
3825
3830
3835
3840
3845
3850
3855
3860
3865
3870
3875
3880
3885
3890
3895
3900
3905
3910
3915
3920
3925
3930
3935
3940
3945
3950
3955
3960
3965
3970
3975
3980
3985
3990
3995
4000
4005
4010
4015
4020
4025
4030
4035
4040
4045
4050
4055
4060
4065
4070
4075
4080
4085
4090
4095
4100
4105
4110
4115
4120
4125
4130
4135
4140
4145
4150
4155
4160
4165
4170
4175
4180
4185
4190
4195
4200
4205
4210
4215
4220
4225
4230
4235
4240
4245
4250
4255
4260
4265
4270
4275
4280
4285
4290
4295
4300
4305
4310
4315
4320
4325
4330
4335
4340
4345
4350
4355
4360
4365
4370
4375
4380
4385
4390
4395
4400
4405
4410
4415
4420
4425
4430
4435
4440
4445
4450
4455
4460
4465
4470
4475
4480
4485
4490
4495
4500
4505
4510
4515
4520
4525
4530
4535
4540
4545
4550
4555
4560
4565
4570
4575
4580
4585
4590
4595
4600
4605
4610
4615
4620
4625
4630
4635
4640
4645
4650
4655
4660
4665
4670
4675
4680
4685
4690
4695
4700
4705
4710
4715
4720
4725
4730
4735
4740
4745
4750
4755
4760
4765
4770
4775
4780
4785
4790
4795
4800
4805
4810
4815
4820
4825
4830
4835
4840
4845
4850
4855
4860
4865
4870
4875
4880
4885
4890
4895
4900
4905
4910
4915
4920
4925
4930
4935
4940
4945
4950
4955
4960
4965
4970
4975
4980
4985
4990
4995
5000
5005
5010
5015
5020
5025
5030
5035
5040
5045
5050
5055
5060
5065
5070
5075
5080
5085
5090
5095
5100
5105
5110
5115
5120
5125
5130
5135
5140
5145
5150
5155
5160
5165
5170
5175
5180
5185
5190
5195
5200
5205
5210
5215
5220
5225
5230
5235
5240
5245
5250
5255
5260
5265
5270
5275
5280
5285
5290
5295
5300
5305
5310
5315
5320
5325
5330
5335
5340
5345
5350
5355
5360
5365
5370
5375
5380
5385
5390
5395
5400
5405
5410
5415
5420
5425
5430
5435
5440
5445
5450
5455
5460
5465
5470
5475
5480
5485
5490
5495
5500
5505
5510
5515
5520
5525
5530
5535
5540
5545
5550
5555
5560
5565
5570
5575
5580
5585
5590
5595
5600
5605
5610
5615
5620
5625
5630
5635
5640
5645
5650
5655
5660
5665
5670
5675
5680
5685
5690
5695
5700
5705
5710
5715
5720
5725
5730
5735
5740
5745
5750
5755
5760
5765
5770
5775
5780
5785
5790
5795
5800
5805
5810
5815
5820
5825
5830
5835
5840
5845
5850
5855
5860
5865
5870
5875
5880
5885
5890
5895
5900
5905
5910
5915
5920
5925
5930
5935
5940
5945
5950
5955
5960
5965
5970
5975
5980
5985
5990
5995
6000
6005
6010
6015
6020
6025
6030
6035
6040
6045
6050
6055
6060
6065
6070
6075
6080
6085
6090
6095
6100
6105
6110
6115
6120
6125
6130
6135
6140
6145
6150
6155
6160
6165
6170
6175
6180
6185
6190
6195
6200
6205
6210
6215
6220
6225
6230
6235
6240
6245
6250
6255
6260
6265
6270
6275
6280
6285
6290
6295
6300
6305
6310
6315
6320
6325
6330
6335
6340
6345
6350
6355
6360
6365
6370
6375
6380
6385
6390
6395
6400
6405
6410
6415
6420
6425
6430
6435
6440
6445
6450
6455
6460
6465
6470
6475
6480
6485
6490
6495
6500
6505
6510
6515
6520
6525
6530
6535
6540
6545
6550
6555
6560
6565
6570
6575
6580
6585
6590
6595
6600
6605
6610
6615
6620
6625
6630
6635
6640
6645
6650
6655
6660
6665
6670
6675
6680
6685
6690
6695
6700
6705
6710
6715
6720
6725
6730
6735
6740
6745
6750
6755
6760
6765
6770
6775
6780
6785
6790
6795
6800
6805
6810
6815
6820
6825
6830
6835
6840
6845
6850
6855
6860
6865
6870
6875
6880
6885
6890
6895
6900
6905
6910
6915
6920
6925
6930
6935
6940
6945
6950
6955
6960
6965
6970
6975
6980
6985
6990
6995
7000
7005
7010
7015
7020
7025
7030
7035
7040
7045
7050
7055
7060
7065
7070
7075
7080
7085
7090
7095
7100
7105
7110
7115
7120
7125
7130
7135
7140
7145
7150
7155
7160
7165
7170
7175
7180
7185
7190
7195
7200
7205
7210
7215
7220
7225
7230
7235
7240
7245
7250
7255
7260
7265
7270
7275
7280
7285
7290
7295
7300
7305
7310
7315
7320
7325
7330
7335
7340
7345
7350
7355
7360
7365
7370
7375
7380
7385
7390
7395
7400
7405
7410
7415
7420
7425
7430
7435
7440
7445
7450
7455
7460
7465
7470
7475
7480
7485
7490
7495
7500
7505
7510
7515
7520
7525
7530
7535
7540
7545
7550
7555
7560
7565
7570
7575
7580
7585
7590
7595
7600
7605
7610
7615
7620
7625
7630
7635
7640
7645
7650
7655
7660
7665
7670
7675
7680
7685
7690
7695
7700
7705
7710
7715
7720
7725
7730
7735
7740
7745
7750
7755
7760
7765
7770
7775
7780
7785
7790
7795
7800
7805
7810
7815
7820
7825
7830
7835
7840
7845
7850
7855
7860
7865
7870
7875
7880
7885
7890
7895
7900
7905
7910
7915
7920
7925
7930
7935
7940
7945
7950
7955
7960
7965
7970
7975
7980
7985
7990
7995
8000
8005
8010
8015
8020
8025
8030
8035
8040
8045
8050
8055
8060
8065
8070
8075
8080
8085
8090
8095
8100
8105
8110
8115
8120
8125
8130
8135
8140
8145
8150
8155
8160
8165
8170
8175
8180
8185
8190
8195
8200
8205
8210
8215
8220
8225
8230
8235
8240
8245
8250
8255
8260
8265
8270
8275
8280
8285
8290
8295
8300
8305
8310
8315
8320
8325
8330
8335
8340
8345
8350
8355
8360
8365
8370
8375
8380
8385
8390
8395
8400
8405
8410
8415
8420
8425
8430
8435
8440
8445
8450
8455
8460
8465
8470
8475
8480
8485
8490
8495
8500
8505
8510
8515
8520
8525
8530
8535
8540
8545
8550
8555
8560
8565
8570
8575
8580
8585
8590
8595
8600
8605
8610
8615
8620
8625
8630
8635
8640
8645
8650
8655
8660
8665
8670
8675
8680
8685
8690
8695
8700
8705
8710
8715
8720
8725
8730
8735
8740
8745
8750
8755
8760
8765
8770
8775
8780
8785
8790
8795
8800
8805
8810
8815
8820
8825
8830
8835
8840
8845
8850
8855
8860
8865
8870
8875
8880
8885
8890
8895
8900
8905
8910
8915
8920
8925
8930
8935
8940
8945
8950
8955
8960
8965
8970
8975
8980
8985
8990
8995
9000
9005
9010
9015
9020
9025
9030
9035
9040
9045
9050
9055
9060
9065
9070
9075
9080
9085
9090
9095
9100
9105
9110
9115
9120
9125
9130
9135
9140
9145
9150
9155
9160
9165
9170
9175
9180
9185
9190
9195
9200
9205
9210
9215
9220
9225
9230
9235
9240
9245
9250
9255
9260
9265
9270
9275
9280
9285
9290
9295
9300
9305
9310
9315
9320
9325
9330
9335
9340
9345
9350
9355
9360
9365
9370
9375
9380
9385
9390
9395
9400
9405
9410
9415
9420
9425
9430
9435
9440
9445
9450
9455
9460
9465
9470
9475
9480
9485
9490
9495
9500
9505
9510
9515
9520
9525
9530
9535
9540
9545
9550
9555
9560
9565
9570
9575
9580
9585
9590
9595
9600
9605
9610
9615
9620
9625
9630
9635
9640
9645
9650
9655
9660
9665
9670
9675
9680
9685
9690
9695
9700
9705
9710
9715
9720
9725
9730
9735
9740
9745
9750
9755
9760
9765
9770
9775
9780
9785
9790
9795
9800
9805
9810
9815
9820
9825
9830
9835
9840
9845
9850
9855
9860
9865
9870
9875
9880
9885
9890
9895
9900
9905
9910
9915
9920
9925
9930
9935
9940
9945
9950
9955
9960
9965
9970
9975
9980
9985
9990
9995
10000
10005
10010
10015
10020
10025
10030
10035
10040
10045
10050
10055
10060
10065
10070
10075
10080
10085
10090
10095
10100
10105
10110
10115
10120
10125
10130
10135
10140
10145
10150
10155
10160
10165
10170
10175
10180
10185
10190
10195
10200
10205
10210
10215
10220
10225
10230
10235
10240
10245
10250
10255
10260
10265
10270
10275
10280
10285
10290
10295
10300
10305
10310
10315
10320
10325
10330
10335
10340
10345
10350
10355
10360
10365
10370
10375
10380
10385
10390
10395
10400
10405
10410
10415
10420
10425
10430
10435
10440
10445
10450
10455
10460
10465
10470
10475
10480
10485
10490
10495
10500
10505
10510
10515
10520
10525
10530
10535
10540
10545
10550
10555
10560
10565
10570
10575
10580
10585
10590
10595
10600
10605
10610
10615
10620
10625
10630
10635
10640
10645
10650
10655
10660
10665
10670
10675
10680
10685
10690
10695
10700
10705
10710
10715
10720
10725
10730
10735
10740
10745
10750
10755
10760
10765
10770
10775
10780
10785
10790
10795
10800
10805
10810
10815
10820
10825
10830
10835
10840
10845
10850
10855
10860
10865
10870
10875
10880
10885
10890
10895
10900
10905
10910
10915
10920
10925
10930
10935
10940
10945
10950
10955
10960
10965
10970
10975
10980
10985
10990
10995
11000
11005
11010
11015
11020
11025
11030
11035
11040
11045
11050
11055
11060
11065
11070
11075
11080
11085
11090
11095
11100
11105
11110
11115
11120
11125
11130
11135
11140
11145
11150
11155
11160
11165
11170
11175
11180
11185
11190
11195
11200
11205
11210
11215
11220
11225
11230
11235
11240
11245
11250
11255
11260
11265
11270
11275
11280
11285
11290
11295
11300
11305
11310
11315
11320
11325
11330
11335
11340
11345
11350
11355
11360
11365
11370
11375
11380
11385
11390
11395
11400
11405
11410
11415
11420
11425
11430
11435
11440
11445
11450
11455
11460
11465
11470
11475
11480
11485
11490
11495
11500
11505
11510
11515
11520
11525
11530
11535
11540
11545
11550
11555
11560
11565
11570
11575
11580
11585
11590
11595
11600
11605
11610
11615
11620
11625
11630
11635
11640
11645
11650
11655
11660
11665
11670
11675
11680
11685
11690
11695
11700
11705
11710
11715
11720
11725
11730
11735
11740
11745
11750
11755
11760
11765
11770
11775
11780
11785
11790
11795
11800
11805
11810
11815
11820
11825
11830
11835
11840
11845
11850
11855
11860
11865
11870
11875
11880
11885
11890
11895
11900
11905
11910
11915
1192

A preferred embodiment of the present invention is shown in FIG. 1. While the present invention may occur within any distributed computer system, the Internet 100 is preferable, because it provides a common platform to access thousands of sources of business data. In particular, Internet 100 provides access to continuously updated business data found on both a company's own site and third party websites. Within Internet 100 are files or documents containing potential business data, indicated by 120A, 120B, 120C and 120D, from for instance corporate websites and third party sources such as press release wire services, newspaper/magazines and professional/trade associations, that contain the name and titles of millions of senior managers in case of the example of compiling a list of senior manager. Sometimes additional contact information such as email address, telephone number or postal address is also included. The present invention is not limited to the type of business data to be extracted or mined.

30

In FIG. 1, Search Engine and/or Spider 130 refers to both i) a specially-developed spider to access information behind a CGI script, accessible only by entering a search string into a text bar or clicking a search button and ii) a publicly-accessible search engine such as Google (www.google.com) or AltaVista (www.altavista.com) which indexes a substantial portion of the web that is accessible through hyperlinks. Searches could be constructed using data from the databases 150A ... 150N. Examples of such databases 150A ... 150N in the context of compiling contact information are, for instance, but not limited to, an organization, job title, and contact database. The organization database contains, for instance, fields for the organization's name, postal address, industry/SIC, third party identifiers (e.g. DUNS #) and primary corporate domain name. However, it is not necessary for all data elements to be available for all records in the organization database. This data is collected by various business data providers, such as MarketModels, Warwick, RI, which sells a product called eBizFile.. The present invention distinguishes general purpose spiders 130 that identify files on, for instance, a site like Yahoo as well as custom build spiders 130 that are content specific and search for content specific information.

A job title database contains, for instance, but is not limited to, common job titles for senior managers of various kinds of organizations. U.S. for-profit corporations (.com or .biz) would commonly use titles such as CEO, Chief Executive Officer, Chairman, Vice President of Marketing, and so on. U.S. universities (.edu) would commonly use titles such as chancellor, provost, dean, professor, and so forth. U.S. non-profit organizations (.org) might use titles such as executive director or managing director. The present invention is not limited to the U.S. and other files or links could be used with different extensions, such as .nl or .uk for websites originating in The Netherlands or the United Kingdom respectively.

A contact database contains previously mined records for senior managers. It contains for instance, but is not limited to, fields for the manager's name, title, company identifier such as DUNS #, manager's postal address, email address, and telephone number, the URL containing the mined information, the date URL was mined, and the date the URL was last updated. A contact database can be partially seeded with existing data from traditional suppliers, such as a manager's name and title.

Databases **150A ... 150N** could be used to form queries to identify the URL containing the desired business data. First, an attempt is made to identify the business data, or in the particular example, contact information within the corporate website. The present invention utilizes different types of information to identify the proper links. Examples of the type of information, in line with the current example, are for instance, but not limited to, corp, corporate, company, team, management, overview, leadership, leaders, staff, exec, executive, bios, biograph, professional, people and profile. The present invention searches also for links based on their proximity on a set of keywords. Examples of such keywords are, for instance, but not limited to, about us, about, analyst info, company, background, contact, us, corporate, officer, executive, management, or combinations of these keywords or the like.

According to the present example, the corporate website is searched to identify a 'management' contact page. This is done by searching for a job title appropriate for the organization, and constraining the search to pages within the corporate site. For example, when trying to find contacts at ABC Corporation, one could use a common title at a for-profit company in combination with the primary domain name of the company. One could also exclude sites with hyperlinks to the abc.com website. Using the AltaVista search engine, the query would be:

(“Chief executive officer”) AND (host:abc.com)

A further refinement would use existing contact information, such as the name of a known contact to help identify the right page. If Joe Smith is the Vice President of Marketing, at 5 ABC Corporation and the goal is to discover the CEO, the query might be refined to be:

“chief executive officer” AND (Smith NEAR Marketing) AND (host:abc.com)

The goal of the effort is to identify a candidate management page on the company’s website 10 and then further parse the page or file to extract and compile business data.

As shown in **FIG. 1**, the text of the candidate page is passed to parser **140**. Parser **140** attempts to identify, normalize and extract business data, or in the particular example contact 15 information. Parser **140** must be able to extract information in a variety of words and formats. For example, the same title may be abbreviated in multiple ways, such as VP or V.P. for Vice President. Numerous other abbreviations are used in listings of senior managers including sr, exec, dir and so on. Email addresses may be explicitly provided or provided through a ‘mail-to’ hyperlink. Telephone numbers may be listed with different formats, such as the full direct dial number, a general company number with an extension or just an 20 extension. Parser **140** translates this semi-structured information into a structured format. Once the title has been extracted and normalized, an attempt is made to find a proper name 25 near (usually immediately preceding) the title. In some cases, for example, the name associated with the title may appear after the title or below or above the title in a webpage. In such cases, for example, the parser **140** first makes a determination of a local content grouping near the title that is associated with that title by using tags and keywords as boundary markers. Once a local content grouping near the title is identified, identification

and extraction of a name, for example, can take place only in the local content grouping.

Parser 140 matches candidate given and surnames to, for instance, a proper name database,

which is, for instance, a database out of the series of databases 150A ... 150N. In cases

where a name does not match any name in a proper name database, certain characteristics of

names, for example, can be used to detect and extract names. Such characteristics that define

names ('inclusive characteristics') or characteristics to discount off names ('exclusion

characteristics'), for instance, are used to extract the words that are names and to discount the

ones that are not. Such inclusion characteristics for a name, for instance, could include

capitalized first letters, consonant and vowel placements and such. The exclusion

characteristics, for instance, but are not limited to, could include presence of apostrophes, or

multiple non-simultaneous capitalized letters, length characteristics and such. The inclusion

characteristics for an email address, for example, but not limited to, could be presence of the

‘@’ sign and the presence of the internet address in the dot-notation format. Additional

inclusion characteristics could include the requirement that the top-level domain name be

from a specific list of top-level domain names. Exclusion characteristics for discounting

words that are not email addresses for example could be absence of an @ sign, or the

presence of a '-' (dash) etc. Once name and title are extracted, a score is computed regarding

the confidence level of the extraction. Presumably, an extraction in which both the job title is

found in the job title database, and the name is found in the proper name word score highly.

Conversely, an extraction in which the job title did not match the job title database, but the

extracted name matched the proper name database, would score lower.

If the contact information has been successfully parsed, the URL for the page with the highest

score (e.g. <http://www.lowellgeneral.org/HTML/Mgmt.html>; in the example provided above

may be added to the manager(s) record in contact database. The URL may be used to directly

access the contact information on subsequent updates. When reaccessed, the date or time

stamp on the page is compared to the last mining date. If the page has been updated since the last mining, contact parsing is repeated.

The preceding contact parser references databases of 'allowable' job title and proper names to extract semi-structured information from HTML webpages. Contact parsing would be considerably simplified should a more formal page markup language, such as XML, come into widespread use. XML allows text on websites to be tagged to, among other advantages, allow easier information extraction.

10 XML is a markup language for documents containing structured information. Structured information contains both content (words, pictures, etc.) and some indication of what role that content plays (for example, content in a section heading has a different meaning from content in a footnote, which means something different than content in a figure caption or content in a database table, etc.). Almost all documents have some structure. A markup language is a mechanism to identify structures in a document. The XML specification defines a standard way to add markup to documents.

15 In HTML, both the tag semantics and the tag set are fixed. An `<h1>` is always a first level heading and the tag `<ati.product.code>` is meaningless. The W3C, in conjunction with browser vendors and the WWW community, is constantly working to extend the definition of 20 HTML to allow new tags to keep pace with changing technology and to bring variations in presentation (stylesheets) to the Web. However, these changes are always rigidly confined by what the browser vendors have implemented and by the fact that backward compatibility is paramount. And for people who want to disseminate information widely, features supported 25 by only the latest releases of Netscape and Internet Explorer are not useful.

XML specifies neither semantics nor a tag set. In fact XML is really a meta-language for describing markup languages. In other words, XML provides a facility to define tags and the structural relationships between them. Since there's no predefined tag set, there can't be any preconceived semantics. All of the semantics of an XML document will either be defined by
5 the applications that process them or by stylesheets.

Referring again to **FIG. 1**, updated business data **160** or contact information is transferred as a

computer product to an interested party or customer **110**, preferably using the Internet **100**.

For example, contact information **160** may be transferred by ftp or email, or they may be
10 located at a secure web site accessed only by those given permission. The interested party
110 may be an end-user (such as the sales & marketing department of a corporation) or a
value-added remarketer (e.g. business data vendors such as Acxiom, infoUSA or Dun &
Bradstreet). In addition, customer **110** receives a computer program product as a result of the
method of the present invention. The computer program product has stored business data
compiled from files located in a distributed computer system. The storage can be in any type
of computer readable medium such as, but is not limited to, recordable type media such as
floppy disks and CD-ROMS and transmission type media such as digital and analog
communication links.

20 Business data **160**, or contact information in the particular example, that is transferred to
interested party **110** may contain varying amounts of information. For example, it may
contain only contact name and title, or it may also include auxiliary contact data (postal
address, email address or telephone number), third party identifiers (such as DUNS #'s) and
time stamps (date of the mining or file creation date of any contact information).

Now addressing **FIG. 2** in combination with **FIG. 1**, **FIG. 2** shows different steps involved in processing or parsing **140** the potential business data according to the present invention.

Once the files or links have been identified in search **130**, each file or link is then parsed **140**.

Once files or links have been identified a list is generated that contains the previously

5 generated files or links for future use in parsing **140**. In parsing **140**, each file is received **200**

and one or more checks **210** are performed. Checks **210** are, for instance, checks whether the

file or link is still available or whether the file or link has moved to a different location. In

case the file or link has moved and the new location is provided, the database that contains a

list of links or files is updated. Various different checks **210** could be performed to ensure the

10 integrity of the files or links before proceeding **212** on the actual processing/parsing in **140**.

Once a file or link has not been approved **214** the program returns **220** to the main program.

Upon acceptance of the file or link an evaluation **230** is performed. Evaluation **230**

15 determines a confidence level of the business data in the file or link. The objective of

evaluation **230** is to determine the likelihood of finding business data and/or the quantity of

business data in that file or link to determine if this file or link is worthwhile to pursue.

Evaluation **230** is done on either a subset of the type of business data under investigation or

on the entire set of business data. The present invention also includes the determination of a

confidence level of finding a set of keywords in the files or links that contain business data.

Different keywords can be used that quickly identify the type of business data available in the

20 files or links. Once a certain threshold is achieved in the evaluation step, the program

continues. Otherwise the program proceeds with the next file or link in the list to be parsed.

The present invention also includes the step of evaluating **a tense 240**, for instance past or

future tense, of the data that is identified. In the example of senior management data it is

25 pertinent that the information be current. In that case, a past tense evaluation is performed

240 to check on past tense relationship of the business data and/or keywords. Once the past

tense relationships are identified, they will be discarded. In another example, it might be important to only select on past tense when a customer **110** wants to get a history on certain events related to their business or marketing.

5 The present invention also includes the step of using one or more tags to identify or extract business data. For instance, but not limited to, tags like **<TR>**, **<TD>**, **<Title>**, **
, **<P>, or **** could be used to identify keywords and/or business data or to establish boundaries of local content grouping to facilitate extracting business data. The present invention is not restricted to the type of tags or equivalent structures in a file or link that can be identified or 10 associated to potential business data. Any other type of system or coded semantics could be used.

15 Using one or more tags, the present invention also includes the step of generating a concordance table **250**. The purpose of table **250** is to associate the business data under consideration and/or keywords with a number or another type of rating system so that the 20 information can be rated and bounded. The rating process is, for instance, but not limited to, the process of going word by word through a file or link and adding to a number of counts when certain keywords and/or tags that are found. An example is, for instance, but not limited to, data containing chairman, CEO, VP etc. that established a count of, for instance, 19, 15, and 12 respectively depending on whether the proper tags and keywords are identified. Each word in the file also, for instance, has a tag and keyword count set associated with it. An example is, for instance, but not limited to, **<P>** Joe Smith **
** Chairman and CEO **
** ABC Corporation **
** John Doe **
** President **<P>**, that established a set of 25 1 **<P>** tag and 0 **
** tags for words Joe and Smith, a set of 1 **<P>** tag and 1 **
** tag for words Chairman and CEO, a set of 1 **<P>** tag and 2 **
** tags for words ABC and Corporation, and a set of 1 **<P>** tag and 4 **
** tags for the words John and Doe. A

bounding process is, for instance, but not limited to, the process where certain localized content groupings around said business data are identified based upon tag and keyword placements. In the example above, the tags **<P>** and the double **

** are, for example, identified as a boundaries around the keywords Chairman and CEO. This bounding process, 5 for example, helps in associating the name Joe Smith with the title Chairman and CEO and helps facilitate the extraction of the correct and associated name later. The name 'John Doe' also for instance, similarly gets associated with the title President and later gets properly extracted in a name detection process. Any database of **150A ... 150N** could be used to deliver the proper keywords as well as the instruction set to identify the proper tags and boundaries. However, the instruction set to identify the proper tags and boundaries could also be part of the program.

10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

The rating in concordance table **250** is beneficial to the step of analyzing **260** from which position in the file and up to which position in the file to do a gross extraction of the business data. Analysis **260** could also be referred to as a break-factor analysis. Important in the analysis **260** is to check the positions of the keywords and tags and identify local word groupings that denote a certain content association with each other. In the example of contact information, it associates the right name with the right title by identifying the boundaries between various names and titles. The analysis includes, but is not limited to, an analysis by identifying the proper keywords and tags starting from the front of a word grouping as well as an analysis by identifying the proper keywords and tags starting from the back of a word grouping. After such an analysis, a number of factors are identified and based on these factors the business data are then ready to be evaluated and extracted **270**. In the analysis step **260** duplicated words are eliminated as well as words that are predefined as disqualifying words. The extraction process **270** includes, but is not restricted to, using databases **150A ... 150N** to identify business data and/or applying inclusion and exclusion characteristics suitable

and appropriate for extracting the required business data. The next step in the parsing **140** method is to normalize the extracted data to a standard format. Such normalization for the contact information example would be to have the names in the first name and last format, and to expand all the acronym'ed titles to their full forms etc. The normalization process, may

5 for example, be used later to identify and eliminate duplicate business data. The next step in the parsing **140** method is to collect the identified, analyzed and selected business data and

group them according to a particular format. Such a format could for instance be, in the example of a list of senior managers, a list that starts with the most senior people first and then going down the list of management seniority such as Chairman, President, CEO, COO,

10 VP etc. Once a set of business data **160** has been compiled it is sometimes necessary to update or refresh the data to keep it current. Therefore the present invention also includes the step of updating **280** the database before it returns **290** back to the main program. Updating

280 is done by using time stamps or date stamps that compare to the last date of mining or the parsing of a particular file or link. Once the stamp on the accessed file is found to be

15 newer/later than the stamp in the compiled data or in the business data database **160**, the method is repeated to update the business data database **160** with the most current information. Date or time stamps are also evaluated for determining whether the business

data in a file is in accordance to the business data that one is looking for. For instance, a press release has a date and time stamp that provides information on whether to proceed with

20 parsing and extracting the business data from that file containing the press release.

It will be clear to one skilled in the art that the above embodiments may be altered in many ways without departing from the scope of the invention. For example, while corporate websites are preferred for obtaining current information, the invention can be extended to extract information from third party websites, such as press release wire services, professional/trade associations or magazine/newspaper websites. This information can be

10
15
20
25

used to augment the contact database if a corporate website does not exist or does not contain a management page. Potential sources may sometimes need to be accessed manually first to learn how access to the contact information is obtained. Typically, business data are either directly accessible by following hyperlinks, or behind a CGI script, accessible only by entering a search string into a text bar or clicking a search button. For hyperlink accessible data, a general purpose search engine such as Google (www.google.com) may be used. For data behind CGI scripts, an automated spider must be tailored to each particular website, because contact information is accessed differently from each web site. For example, sources of contact information that can be accessed by a hyperlink are useful. Conversely, Bizjournal.com provides a full-text searchable archive of numerous local business newspapers. Both types of sites (hyperlink and script) may be accessed using an automated spider.

In addition, those skilled in the art will appreciate that the method and apparatus of the present invention, as well as the product resulting from this, is capable of being distributed in the form of a computer readable medium of instructions in a variety of forms. The present invention applies equally regardless of the particular type of signal bearing medium used to actually carry out the distribution. Examples of computer readable medium include: recordable type media such as floppy disks and CD-ROMS and transmission type media such as digital and analog communication links. In addition, the present invention could be implemented and coded in different programming languages such as, but not limited to, for example C and C++ programming languages, JAVA or Java script, or DHTML.

The present invention has now been described in accordance with several exemplary embodiments, which are intended to be illustrative in all aspects, rather than restrictive. Thus, the present invention is capable of many variations in detailed implementation, which may be